『日本語歴史コーパス　鎌倉時代編Ⅰ説話・随筆』形態論情報の概要

2016 年 3 月 31 日

池上　尚

１．２種類の言語単位

（1）用例収集を目的とした**短単位**
（2）言語的特徴の解明を目的とした**長単位**

『日本語歴史コーパス　鎌倉時代編Ⅰ説話・随筆』で採用したこの 2 種類の言語単位は、『現代日本語書き言葉均衡コーパス(BCCWJ)』で採用した単位を基に設計したものである。基となっている『BCCWJ』の言語単位は『日本語話し言葉コーパス（CSJ)』との互換性の保持を図り、国立国語研究所が行った語彙調査の単位を基に設計された。

本コーパスの言語単位は、通時的な日本語研究で利用するために、現代語のコーパスとの互換性の保持を図っている。これまでに国立国語研究所が実施してきた語彙調査における言語単位のうち、短い単位の系列に属するものが「短単位」、長い単位の系列に属するものが「長単位」である。なお、長単位・短単位認定規程は、『BCCWJ』の規程をそのまま用いるのではなく、本コーパス用に修正・拡張を行っている。

短単位・長単位とも、代表形（語彙素読み）・代表表記（語彙素）・品詞・活用型・活用形を与える。代表形は国語辞典の見出しに、代表表記はその見出しに与えた漢字等の表記に相当するものである。

２．短単位の概要

短単位は、言語の形態的側面に着目して規定した言語単位である。短単位の認定にあたっては、まず意味を持つ最小の単位（最小単位）を規定し、その最小単位を文節の範囲内で短単位認定規程に基づいて結合させる（もしくは結合させない）ことで認定する。

（１）最小単位

- 最小単位は現代語において意味を持つ最小の単位である。本コーパスにおける最小単位については、現代語との関連を重視して、原則として現代語を対象とした最小単位認定を行うが、必要に応じて、使用実態に基づき個別の判断をすることがある。語種等により、次のように認定する。

  ※「／」は最小単位の分割位置を表す。

和語　　：花／は／さかり／に／、／月／は／くま／なき／を／のみ／見る／もの／か／は／
漢語　　：／関／白／　／大／納／言／　／祈／請／
外来語　：／菩薩／　／瑠璃／　／阿闍梨／
記号　　：／。／　／・／
人名　　：／平／将門／　／白／楽天／
地名　　：／大和／の／国／宇陀／の／郡

- 上記のように認定した最小単位を、短単位認定のために下表のとおりに分類する。

表１　最小単位の分類

分類		例
一般		和語　：春　花　あはれ　言ふ　言葉　…
		漢語　：関　白　加　持　…
		外来語：阿闍梨　菩薩　瑠璃　…
付属要素		接頭的要素：相　御（おおん、ご、み）　打ち　なま　…
		接尾的要素：君（ぎみ）　難し　気（げ）　様（さま）　…
その他	記号	、　・　。　「　」　…
	数	一　二　十　百　千　幾　数　何　…
	固有名	人名：源　貫之　行基　くうすけ　…
		地名：大和　土佐　住吉　吉野　逢坂　鞍馬　…
	助詞・助動詞	の　を　ぞ　こそ　し　る・らる　ず　まじ　まほし　なり　…

（２）短単位

- 短単位の認定規定は、上表の分類ごとに適用すべき規定が定められている。その規定に基づき、最小単位を結合させる（又は結合させない）ことによって、短単位を認定する。以下、「一般」・「数」・その他に分けて、短単位認定規定の概要を示す。
  ※「｜」は短単位の分割位置を、「=」は短単位を切らないことを示す。

［１］一般

《和語・漢語》

最小単位 2 つの結合までを 1 短単位とする。

【例】｜山｜　｜里｜　｜山=里｜　｜ならび=なし｜　｜心=のどか｜
｜法=師｜　｜右｜大=将｜

例外：複合動詞は原則として分割する。
【例】｜聞き｜渡る｜　｜出で｜来｜

例外：切る位置が明確でないもの、あるいは切った場合と一まとめにした場合とで意味にずれがあるものは、3 最小単位以上の結合であっても 1 短単位とする。
【例】｜大殿籠もる｜　｜観世音｜

例外：最小単位が 3 つ以上並列した場合、それぞれの最小単位を 1 短単位とする。
【例】｜銭｜絹｜布｜綿｜　｜馬｜鞍｜牛｜車｜

《外来語》
1 最小単位を 1 短単位とする。
【例】｜紺｜瑠璃｜　｜菩提｜講｜

［２］数
「数」以外の最小単位と結合させない。「数」どうしの結合は、一・十・百・千の桁ごとに 1 短単位とする。「万」「億」等は、単独で 1 短単位とする。
【例】｜二十｜四｜日｜　｜十｜万｜億｜　｜二三十｜束｜

［３］その他
1 最小単位を 1 短単位とする。

付属要素	｜相｜見る｜　｜者｜ども｜　｜堪｜がたし｜
助詞・助動詞	｜夜中｜ばかり｜に｜や｜なり｜ぬ｜らん｜　｜そ｜の｜
記号	｜、｜　｜「｜
人名	｜平｜将門｜　｜白｜楽天｜　｜恵心｜僧都｜
地名	｜大和｜の｜国｜宇陀｜の｜郡｜　｜吉野｜山｜

- 短単位データの作成は自動形態素解析によって行われている。形態素解析処理は形態素解析器に「MeCab」、解析用辞書に「中古和文 UniDic」を使用している。ただし、『今昔物語集（本朝部）』の非コアデータについては、コアデータを学習用コーパスとして作成した「和漢混淆文 UniDic」によって再解析を施している。非コアデータの詳細については、池上ほか（2015）を参照されたい。

３．長単位の概要

長単位は、言語の構文的な機能に着目して規定した言語単位である。長単位の認定は、文節の認定を行った上で、各文節の内部を規定に従って自立語部分と付属語部分とに分割していくという手順で行う。

（１）文節

- 長単位の認定にあたっては、まず文節の認定を行う。現代語の文節は、一般に付属語又は付属語連続の後ろで切れる。このほかに、本コーパスでは、付属語を伴わない自立語であっても、主語・主題、連用修飾、連体修飾の各成分の後ろで切るといった規定を設けた。
- 文節を認定する上で問題となることの一つに、固有名、「―が～」「―つ～」「―の～」で１短単位と認める体言句、副助詞が挿入された複合動詞がある。これらについては、内部にある付属語の後ろでは切らないこととする。
- 複合辞は付属語として認めない。

※「｜」は文節の分割位置を、「=」は文節を切らないことを表す。

｜小野小町｜　｜物の具｜　｜雁が音｜　｜滝つ瀬｜　｜ありのまま｜　｜権守（ごんのかみ）｜
｜北の方｜　｜取り=も=あへず｜　｜思ひ=ぞ=返す｜

（２）長単位

- 長単位は、上記の文節を規定に基づいて分割する（又は分割しない）ことによって認定する。文節を超えることはない。以下、長単位認定規定の概要を示す。

※「｜」は長単位の分割位置を、「‖」は注目している長単位の分割位置を、「=」は長単位を切らないことを示す。

［１］記号は１長単位とする。

【例】｜「｜奥山｜に｜、｜猫また｜と｜いふ｜もの｜あり｜て｜、｜人｜を｜食ふ
｜なる｜」｜と｜、｜人｜の｜言ひ｜ける｜に｜、｜

［２］付属語は１長単位とする。

【例】｜「｜奥山｜に｜、｜猫また｜と｜いふ｜もの｜あり｜て｜、｜人｜を｜食ふ
｜なる｜」｜と｜、｜人｜の｜言ひ｜ける｜に｜、｜

［３］主語・主題、連用修飾成分、連体修飾成分の後ろで切る。

【例】｜あはれ｜なる｜こと｜多かり｜。｜
｜智恵｜無き｜者｜は｜此く｜謀るる｜也｜。｜

［４］体言に形式的な意味の「す」「きこゆ」「はべり」「まゐる」「つかうまつる」が直接続く場合、切り離さない。

【例】｜物語=する｜に｜　｜安置=し給へ｜り｜

［５］「御（おほん・お・み・ご）～す・きこゆ」「～おはす・おはします・きこゆ・さぶらふ・たてまつる・たまふ・つかうまつる・はべり・もうす」という形式の敬語表現は、全体を 1 長単位とする。

【例】｜御曹司=し｜て｜　｜出し進り給へ｜　｜返し給ひ候は｜ん｜

上記形式中に付属語が含まれる場合、切り離さない。

【例】｜御覧じ=興ぜ=させ=給ひ｜ける｜

［６］同格の関係にある体言連続は切り離さない。

【例】｜父=三位｜　｜薩摩守=忠度｜

［７］並列された語は切り離さない。

【例】｜形=有様｜　｜道俗=男女｜

［８］係り受けを重視し、付属語を切り出すのは不適切なものを連語として認める。

【例】｜知ら=ず=顔｜　｜思ひ=の=ほか｜

- 長単位データの作成は、人手修正済み短単位データを基に、長単位解析器 Comainu によって長単位の自動構成を行っている。

４．品詞付与方針

（１）短単位・長単位の相違点

短単位と長単位の品詞体系は共通であるが、品詞付与方針が異なる。短単位では可能性を考慮した品詞を付与しており、「名詞-普通名詞-形状詞可能」等がある。これに対して長単位では文脈に即して品詞を付与する方針をとり、名詞-普通名詞-〇〇可能といった品詞は設けない。例えば、「哀れ」は短単位では「名詞-普通名詞-形状詞可能」であるが、長単位では文脈に則し「もののあはれを知らざりけり」の場合は名詞を、「物のあはれなる夕暮の空」の場合は形状詞を付与する。

（2）他のコーパスと異なる特殊な処理

本コーパスでは、他のコーパスと異なる処理を施した箇所が少なからずある。全体に関わる特に注意すべき特殊な品詞付与例について以下に示す。『今昔物語集（本朝部）』における処理の詳細については、冨士池ほか（2013）を参照されたい。

表２　『日本語歴史コーパス　鎌倉時代編Ⅰ説話・随筆』の特殊な品詞

品詞	内容	例
解釈不明	解釈不明の箇所	｜みな｜けいし｜ぬれ｜ば｜（語義不明） ｜思え｜つれる｜に｜（ツレバとツルニの混態） ｜口覆ひ｜し｜う｜（テの誤写か）
漢文	訓点のない漢文	｜一伏三仰不来待書暗降雨恋筒寝｜
題	『今昔物語集（本朝部）』における、説話冒頭の題	｜越後国神融聖人縛雷起塔語第一｜
意識的欠字（一般）	『今昔物語集（本朝部）』における、具体表記を保留した欠字	｜七条｜より｜は｜␣｜（方位の明記を期した意識的欠字）
意識的欠字（人名−一般）		｜横川｜に｜⇦␣⇨｜と云ひ｜て｜道心｜有る｜聖人｜有り｜
意識的欠字（人名−姓）		｜但馬｜の｜前司｜⇦␣⇨｜千包｜と｜云ふ｜人｜の｜
意識的欠字（人名−名）		｜中臣｜の｜⇦␣⇨｜と｜云ふ｜者｜有｜けり｜
意識的欠字（地名）		｜信濃｜の｜国｜⇦␣⇨｜と｜云｜所｜に｜
意識的欠字（数詞）		｜三｜月｜␣｜日｜の｜事｜也｜
意識的欠字（漢字表記保留）	『今昔物語集（本朝部）』における、漢字表記を期した意識的欠字	｜先年｜の｜御｜⇦␣⇨｜の｜喜く｜候｜しか｜ば｜
破損	『今昔物語集（本朝部）』における、破損による欠字	｜其｜の｜形｜端厳｜なる｜⇦␣⇨｜比なし｜。｜
欠損	原文欠損箇所	｜此｜に｜依｜て｜人｜皆｜（以下欠）｜

参考文献

池上　尚・鴻野知暁・河瀬彰宏・片山久留美（2015）「『今昔物語集』のコーパス化における非コアデータの精度向上作業」『第 8 回コーパス日本語学ワークショップ予稿集』pp.65-74

小椋秀樹・須永哲矢（2012）『中古和文 UniDic 短単位規程集　平成 21（2009）-平成 23（2011）年度科研費補助金　基盤研究（C）「和文系資料を対象とした形態素解析辞書の開発」研究成果報告書 2』

国立国語研究所コーパス開発センター（池上　尚）編（2016）『『日本語歴史コーパス　平安時代編』形態論情報規程集』大学共同利用機関法人人間文化研究機構国立国語研究所コーパス開発センター

冨士池優美（2012）「中古和文における長単位の概要」『第 2 回コーパス日本語学ワークショップ予稿集』pp.51-58

冨士池優美（2015）「『日本語歴史コーパス　平安時代編』の形態論情報」『コーパスと日本語史研究』ひつじ書房 pp.237-280

冨士池優美・河瀬彰宏・野田高広・岩崎瑠莉恵（2013）「『今昔物語集』のテキスト整形」『第 4 回コーパス日本語学ワークショップ予稿集』pp.125-134

参考 URL

「中古和文 UniDic」http://www2.ninjal.ac.jp/lrc/index.php?UniDic

「MeCab」http://code.google.com/p/mecab/

「Comainu for 中古和文」https://osdn.jp/projects/comainu-emj/